OraGIS and Loom: Spatial and temporal extensions to the ORA Analysis Platform

George B. Davis, Jamie Olson, and Kathleen M. Carley

June 2008
CMU-ISR-08-121

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Increasingly, data available to network analysts includes not only relationships between actors but measurements of entity attributes and relations through time and space. Integrating this information with existing dynamic network analysis techniques demands new models and tools. This paper introduces two extensions to the ORA dynamic network analysis platform intended to meet this need. The first, OraGIS, provides geospatial visualization and clustering algorithms. The second, Loom, assists in the analysis of agent movements through a discrete state space (such as a set of named locations) over time. We discuss the capabilities of both tools and their integration with the traditional analytics in the ORA platform.

This work was supported in part by the National Science Foundation under the IGERT program (DGE- 9972762) for training and research in CASOS, and the Office of Naval Research under Dynamic Network Analysis program (N00014-02-1-0973, ONR N00014-06-1-0921, ONR N00014-06-1-0104). Additional support was provided by CASOS - the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
OraGIS and Loom: Spatial and temporal extensions to the ORA Analysis Platform	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Increasingly, data available to network analysts includes not only relationships between actors but measurements of entity attributes and relations through time and space. Integrating this information with existing dynamic network analysis techniques demands new models and tools. This paper introduces two extensions to the ORA dynamic network analysis platform intended to meet this need. The first, OraGIS, provides geospatial visualization and clustering algorithms. The second, Loom, assists in the analysis of agent movements through a discrete state space (such as a set of named locations) over time. We discuss the capabilities of both tools and their integration with the traditional analytics in the ORA platform.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**17**	

Keywords: network analysis, GIS, geospatial, time series

1 Introduction

Traditional social network analysis (SNA) focuses on measurement and analysis of relationships between a single set of agents [9]. In contrast, dynamic network analysis (DNA) is characterized by opportunistic use of available data to model all visible aspects of a system [1]. Typically, this includes multiple node types, node attributes, and snapshots of relationship structure over time [1]. The ORA analysis platform [2] consists of tightly integrated tools and data formats designed to allow the analysis and representation of data with all these aspects.

Recent proliferation of sensor systems has produced many datasets which feature geospatial and temporal information about agent activities in addition to the attributes and relationships typically measured. Examples include data from GPS sensors embedded in vehicles or devices, logs of online activities, and collected data from intelligence networks. In this paper we discuss new features in the ORA analysis platform targeted at this type of data, using as an example a real dataset drawn from sensors placed in merchant marine shipping vessels. We guide the reader through a simple analysis of this data which utilizes the new capabilities within ORA.

The rest of the paper is organized as follows. In section 2, we discuss the representations of this type of data by describing new extensions to our DyNetML markup language to include location and time series information. Section 3 discusses the need for distinct visualizations of time and space data, and introduces the OraGIS and Loom tools which provide these. Section 4 discusses the analytic methods provided by these tools to assist sense making. Finally, we conclude in section 5 with a summary of capabilities and a discussion of future work.

2 Representation

ORA is capable of importing and exporting a variety of formats [2], but its native representation is the format DyNetML [5]. Because DyNetML is XML based, it can be easily parsed, modified and generated by a variety of publicly available tools. Using also XML provides a good combination of machine readability and human comprehensibility.

OraGIS and Loom both use extensions to DyNetML as the persistent representation of their respective data types. Figure 1 reproduces the structural diagram of DyNetML given by [5], revised to include the new extensions (marked in grey).

2.1 OraGIS

OraGIS can read either full DyNetML files (IE those consisting of a DynamicNetwork element) or fragments rooted at the TrailSet element. In either case, the file must define at least one Nodeset of type Location, with at least one Location node having identifying geospatial information. Sample

[1] Subsets of these augmentations have been discussed within SNA literature. DNA can be thought of as the explicit study of new challenges arising from modeling all simultaneously.

[2] As of this printing, these include CSV, TDF, SQL based databases, Analyst Notebook, UCINet, GraphML, and PAJEK.

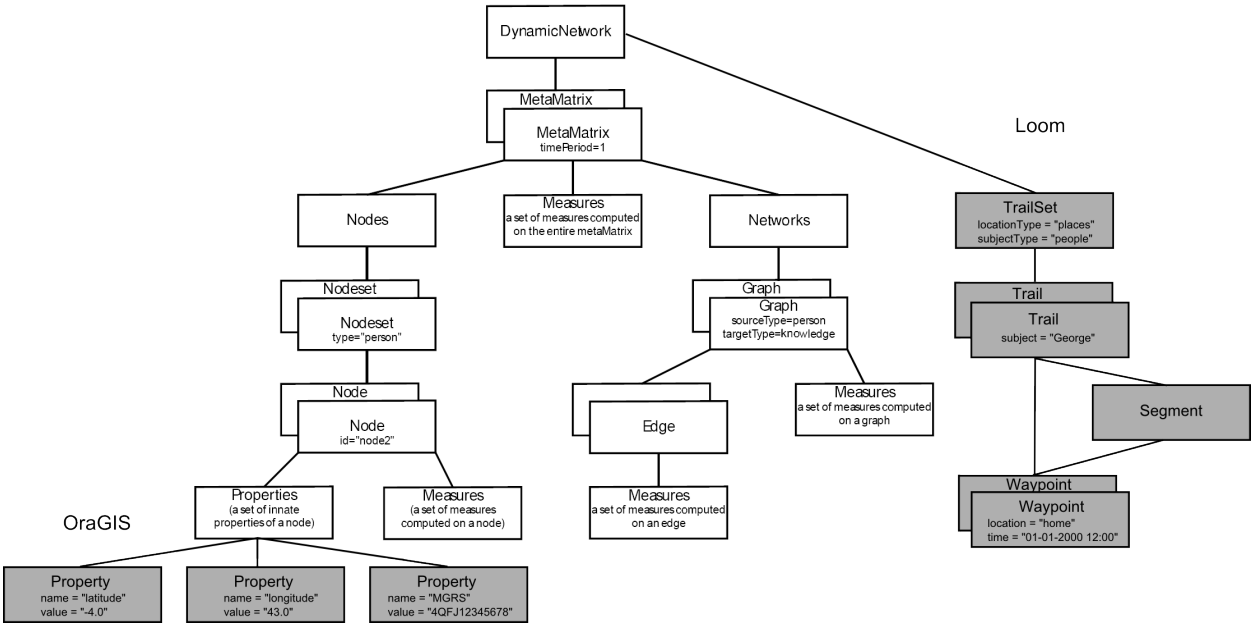

Figure 1: DyNetML Extensions for OraGIS and Loom

code of a compliant NodeSet is shown in figure 2. Location information can be specified either by using decimal latitude and longitude attributes or with a single MGRS[6] attribute.

2.2 Loom

Previous versions of DyNetML have used a "time" attribute at the level of the MetaMatrix element to allow the user to distinguish multiple MetaMatrices acting as snapshots of the same nodesets and relations at various points. This representation is general, in the sense that it can potentially represent any progression of MetaMatrix states, including the entrance and exits of nodes or entire nodesets and relations. However, it encounters problems when encoding high frequency data. Since the entire metamatrix must be reproduced for each time point, even a modestly sized network can only be sampled at a few intervals before exceeding the typical desktop computer's memory constraints. This limitation is particularly unfortunate in cases where data is "temporally sparse", *i.e.* only a few network aspects change between observations. Data of this type can be efficiently encoded by recording only changes in a network rather than entire snapshots, an approach we refer to as *differencing*.

Loom extends DyNetML to permit a differencing representation of temporal data of a particular type: the progression of partners taken on by a subject node that an carry only one instance of a relation that at any given time. Restricting ourselves to changing affiliation data allows us to create a representation that is as efficient as possible, in terms of both space required and computation time for certain analyses. Changing relations of this kind are often encountered in data measuring containment (i.e. changing nationalities of a disputed city) or affiliation (i.e. Agent X Employer). However, for the rest of the paper we discuss loom in terms of it primary purpose, analyzing the

```
<?xml version="1.0" standalone="yes"?>
<DynamicNetwork>
    <MetaMatrix id="Cities">
        <documents></documents>
        <nodes>
            <nodeset type="Location" id="Location">
                <node id="_40">
                    <properties>
                        <property name="name" type="string"
                                value="40"/>
                        <property name="longitude" type="float"
                                value="-99.811409"/>
                        <property name="latitude" type="float"
                                value="32.401025"/></properties>
                </node>
                <node id="_60">
                    <properties>
                        <property name="name" type="string"
                                value="60"/>
                        <property name="longitude" type="float"
                                value="-67.12814"/>
                        <property name="latitude" type="float"
                               value="18.406152"/></properties>
                </node>
                ...
                ..
                .
```

Figure 2: Example of an OraGIS-compliant nodeset in a DyNetML file

path of an agent between physical locations. Our representation of this type of data is accomplished by the following new MetaMatrix entities.

- A *waypoint* is a triplet binding a subject to a location at a specific time. Is is encoded via

  ```
  <Waypoint time="yyyy-MM-dd hh:mm:ss" location="location_id"/>
  ```

- A *segment* combines two waypoints into a single entity, to permit designation of properties that are specific to the transition between waypoints rather than the points themselves. It is encoded as

  ```
  <Segment><Waypoint .../><Waypoint .../></Segment>
  ```

- A *trail* contains both segments and free-floating waypoints recording the progression a single subject node. It is encoded as

  ```
  <Trail subject="subject_id">
  <Waypoint .../>
  <Segment>.../Segment>
  ...
  </Trail>
  ```

- The *trailset* gathers all trails stemming from the same relation or data source. It defines a location nodeset, containing potential nodes for the "one" side of the one-to-many relation, and a subject nodeset which can affiliate with only one location at a time. It is encoded as

  ```
  <TrailSet id="id" locationType="locSet" subjectType="subjSet">
  <Trail ...>...</Trail>
  ...
  </TrailSet>
  ```

The series of waypoints and segments within a trail must follow certain rules which cannot be formally stated in an XML schema, but are enforced when parsing into ORA. All time periods on waypoints must be strictly sequential, with one exception: the first waypoint of one segment may have identical time and location as the last waypoint of the previous segment. This allows encoding continuous data about a subject so that the location is never unknown. If the identical waypoints have associated attributes are measures, they will be unioned when the file is parsed into memory.

The trailset element is placed on the same level as the metamatrix, because it is not bound to the specific time associated with a metamatrix snapshot. Like other metamatrix entities, each of the above can be annotated with user-defined properties to record all data known regarding each element. Although ORA defines some conventions – such as the location encodings used by OraGIS – it is up to the user to standardize his or her use of entity properties to aide analysis.

3 Visualization

3.1 OraGIS

OraGIS allows the integration of geospatial information into the analysis of relational data. Although it may be tempting to treat geospatial locations as simply yet another node attribute, that would dramatically under-emphasize the importance of spatial distance and proximity in a number of problems relevant for DNA. Geospatial location, like temporal information, is a privileged dimension. Spatial proximity is a fundamental fact of the physical world and it would be misguided to ignore that fact. This is especially evident in such contexts as epidemiology, where physical interaction is required for the transmission of disease, or shipping networks, where collocation is required for goods to be transferred.

Once that is acknowledged, we face the question of how best to incorporate geospatial information into relational analysis. OraGIS supports many features that are common in traditional geospatial analysis such as zoom, pan and select. Figure 3 shows the OraGIS user interface. Most of these translate intuitively from the geospatial domain into the network domain (e.g. zoo, selection(figure 5)). These are augmented by a variety of common network analysis methods. In particular, OraGIS allows analysts to adjust the visual features (e.g. color, size) of places according to network properties (see figure 6). Users also have the ability to perform further analysis on selected areas by saving a selection as a new new metamatrix (see figure 5). In doing this, the full capabilities of the Ora suite can be leveraged.

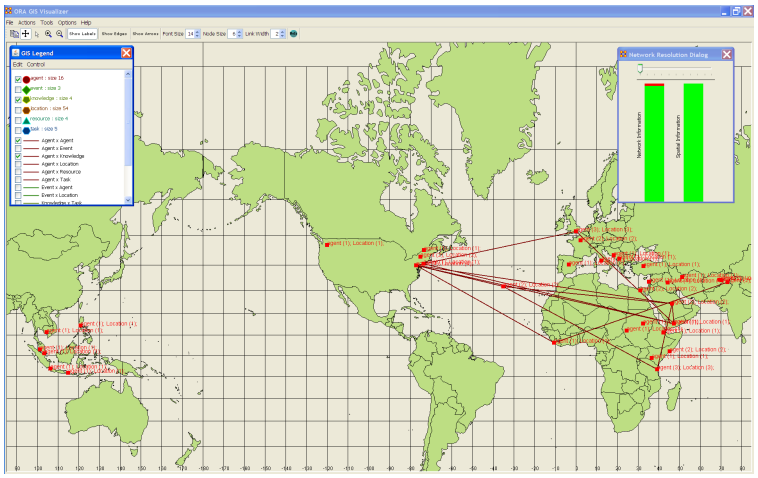

Figure 3: OraGIS user interface

Just as many standard techniques for geospatial analysis can be leveraged in geospatial DNA, several potential pitfalls of geospatial analysis also appear in geospatial network analysis. In particular, the modifiable area unit problem is of particular relevance [8]. The modifiable area unit problem is the danger that the results of a particular analysis are not representative of the actual data but simply an artifact of the aggregation and segregation of geospace into comparable units.

In discussing the aggregation and segregation of geospace, it becomes necessary to differen-

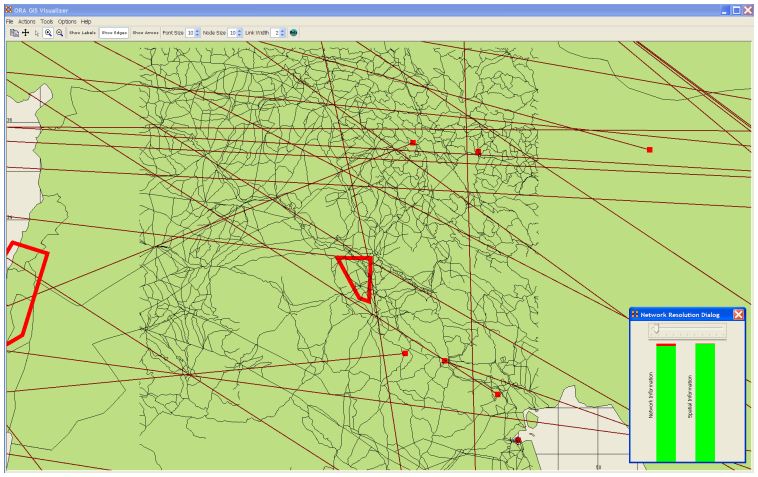

Figure 4: Roads shapefileoverlaidd with network visualization

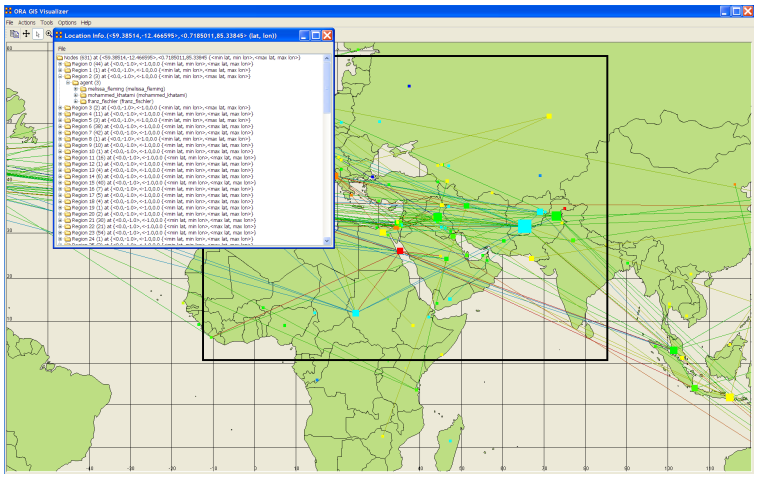

Figure 5: Selecting a region for analysis

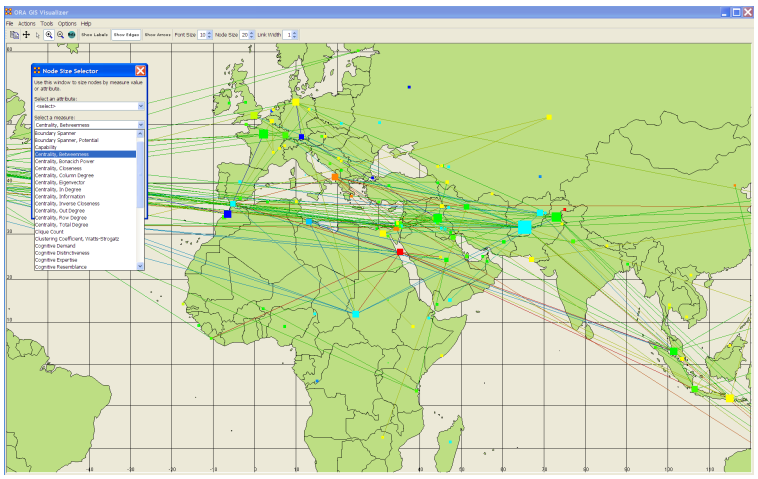

Figure 6: Places colored according to a Newman-Girvan grouping and sized according to betweenness centrality

tiate between *locations* and *places*. We refer to *locations* as precise positions in geospace, most commonly a <latitude, longitude> pair. In contrast, we use *places* to refer to the meaningful regions in which a research question is posed. For example, a social network may contain individuals labeled with home addresses *locations*. We may use this dataset to determine the US cities that are most connected in this social network. In this context, we would use US cities as *places* for the analysis.

This distinction is important in discussing geospatial network analysis. The domain of geospace is continuous and as such, geospatial data is likely to be continuous. The relational data underlying network analysis, is defined by connections between discrete entities and/or attributes. In order to include geospatial places in network analysis it is necessary to aggregate continuous locations into meaningful places.

The importance of aggregation for geospatial network analysis means that any system for geospatial DNA must have some method of assessing the modifiable areal unit. OraGIS uses user-adjustable geospatial clustering [4] to dynamically aggregate locations into places. This allows analysts to not only select their perceived appropriate level of analysis, but also to easily perform sensitivity analysis by increasing and decreasing the resolution level.

In addition to sensitivity analysis, OraGIS also provides a quantitative measure of information loss due to geospatial aggregation of the network[7]. This is presented as the proportion of network's information content that is preserved in the current aggregation. By combining sensitivity analysis with the information loss metric, OraGIS helps analysts to make more informed decisions about the most appropriate level of analysis.

3.2 Loom

Loom uses a separate visualization to render one trailset at a time, outside the context of the full metamatrix. Figure 8 shows the waterfall diagram used by Loom to visualize changing relations

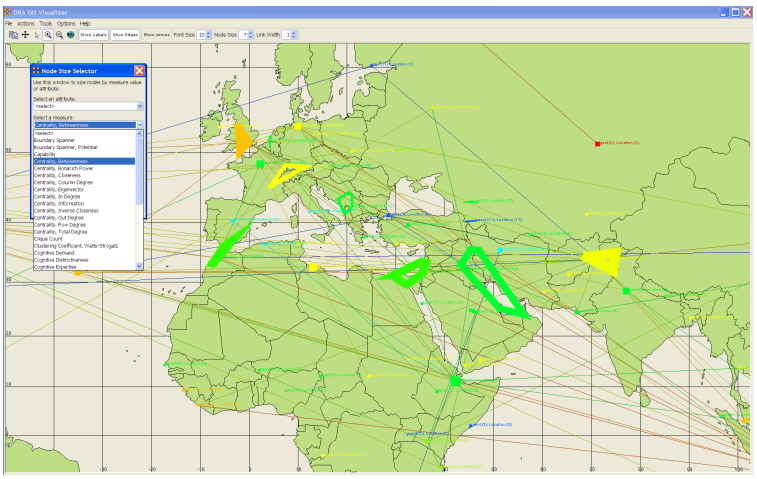

Figure 7: Clustered places colored according to a Newman-Girvan grouping and sized according to betweenness centrality

over time.

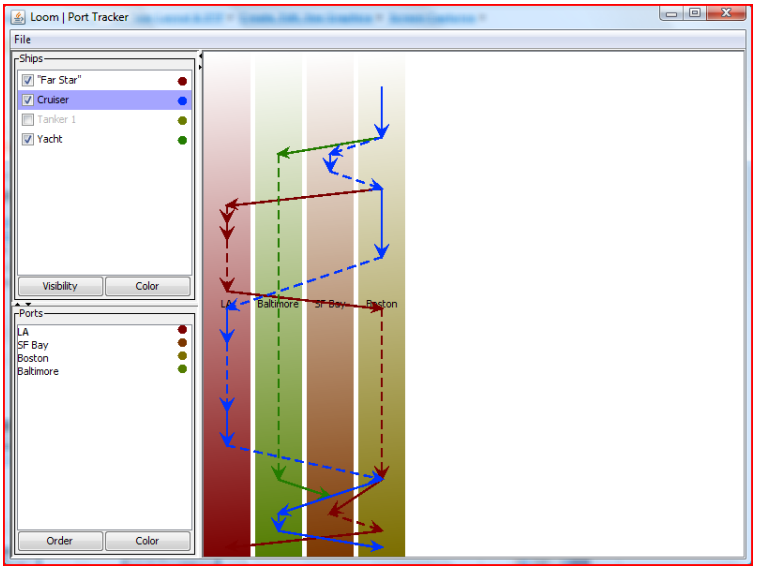

Figure 8: Loom screen featuring waterfall chart.

Locations are shown as vertical strips, laid side by side across the viewing area. The series of transitions taken by a subject node are displayed as a series of arrows indicating transitions between location associations. When multiple subjects are viewed at once, each is given a unique color to distinguish trip segments. The vertical axis encodes time, with the top of the screen being earliest and the bottom latest. Line endpoints correspond to waypoints in the DyNetML representation. When waypoints are joined within a segment, a solid line is used to show that the transition was explicitly observed. When waypoints are free floating, a dashed line is used to show that the transi-

tion was inferred based on observed waypoints. If the source data is potentially missingtransitionss (as with most sensor data), the viewer should consider the possibility that additional movements took place in the time period spanned by the dashed line.

We intended to make several visual comprehensions easy for the user. At the simplest level, we wanted to enable first order queries of the type "what locations are associated with which subjects" and vice versa. We also wanted to enable exploration of the attributes of those relations. Figure 9 shows a submenu expansion accessed by right clicking on a transition. The submenus act as a simple browser, giving quick reference to the precise times and attributes of subjects, locations, segments and waypoints.

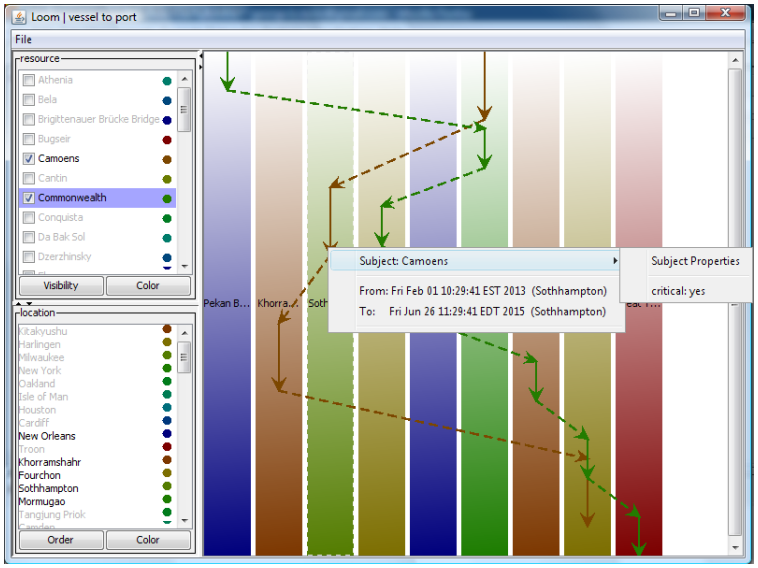

Figure 9: Exploring transition and entity properties via Loom submenus.

We also wanted to elucidate some more complex patterns. One example is cyclical behavior, such as an agent switching locations on a regular schedule. Figure 10 shows an example of a two schedules from real world data featuring periodic patterns due to transport routines. On the left is an extremely simple routine in which a transport vessel has reported a regular schedule of movements between two anchorages. However, the dashed arrow near the beginning indicates an inferred transition that fits the pattern but was not directly reported in data – an exception that could potentially raise a flag for further investigation. On the right is a much more complex trail that nonetheless demonstrates recurrence of certain routines, such as the trip from location "MKP" to "RFP" via "RFT4" and "SEAS". The additional stop at "RFT1" in the middle trip could indicate anexceptionall event.

Another analysis we wanted to make visually easy was determining whether two subjects had shared a location during the same time period. Figure 11 shows three subjects which visit the same locations, but usually at different times. If this data were collapsed into a single network showing all locations visited by the two subjects within the time period (a commonly used technique for "chunking" data on changing networks), the false impression would be given that all three subject

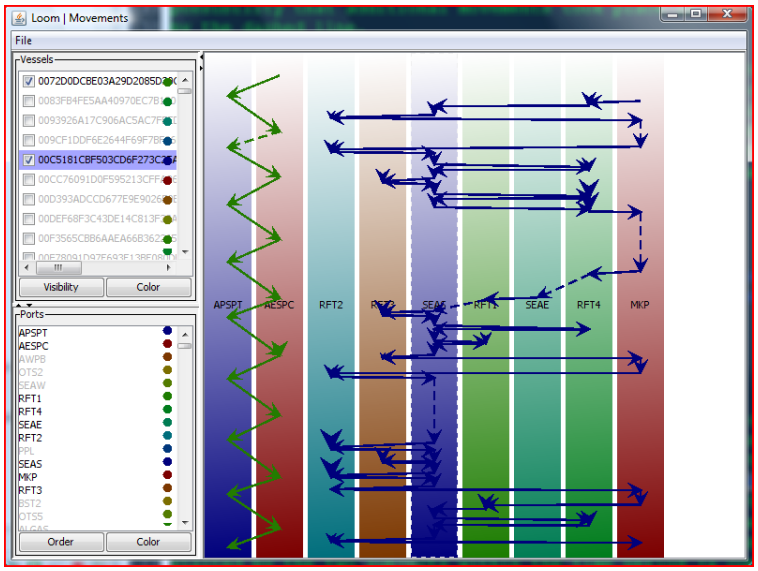

Figure 10: Loom visualization demonstrating transportation routines.

collocated during the time period. By scanning each subject line in Loom, we can see that the red subject had briefcollocationsns with both other actors, but that the blue and green actors never met.

4 Analysis

So far we have discussed Loom and OraGIS in isolation. In this section, we will demonstrate that they can be integrated with each other and with the rest of ORA into a comprehensive analysis workflow. The dataset we will be using is a log of AIS transmissions containing the identity and locations of merchant marine vessels in the English Channel over a five day period. Figure 12 shows the trails in geographic context.

Although our AIS data is easily representable as a trailset, it is not easily analyzed within Loom. This is because ships report their locations as continuous GPS coordinates. Since every coordinate is interpreted as a distinct location, the trails are degenerate in the sense that no two ships visit the same location or revisit their own path. To gain insight from a Loom analysis, we must merge nearby points into a smaller set of interesting locations which will be revisited and shared between ships. Figure 13 shows the results of a clustering of points based on geospatial density. The regions identified correspond to major ports and shipping lanes in the dataset.

We can generate a new trailset in which the old trails are projected onto these clustered locations by selecting `File->Save->As Trailset` in OraGIS. Figure 12 shows several trails from this new set visualized by Loom. These trails now demonstrate the patterns we identified in the previous section. For example, two ships clearly colocate at region 9.

Loom can project it's temporal model back into a MetaMatrix for analysis in ORA. By selecting `File -> Export DyNetML...` we can save a new DyNetML file with two networks:

- A "visit" matrix which simply records which ships visited which ports. Visualizing this

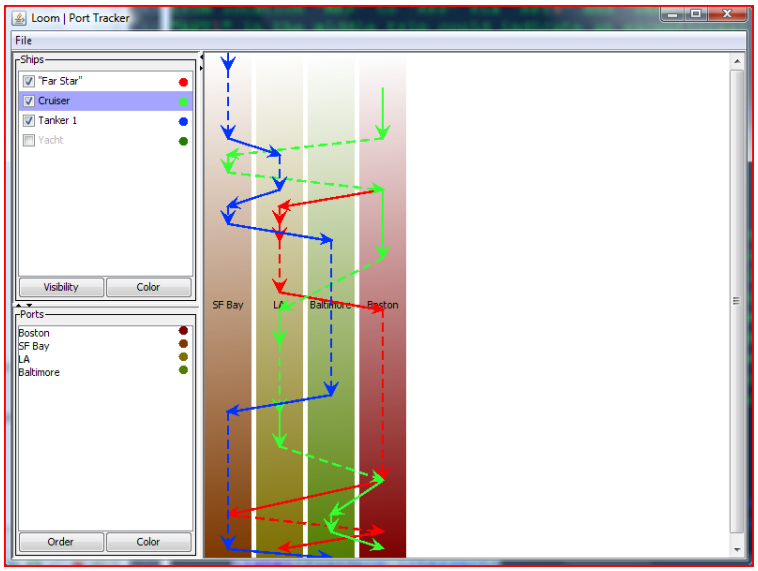

Figure 11: Detecting subject meetings with Loom.

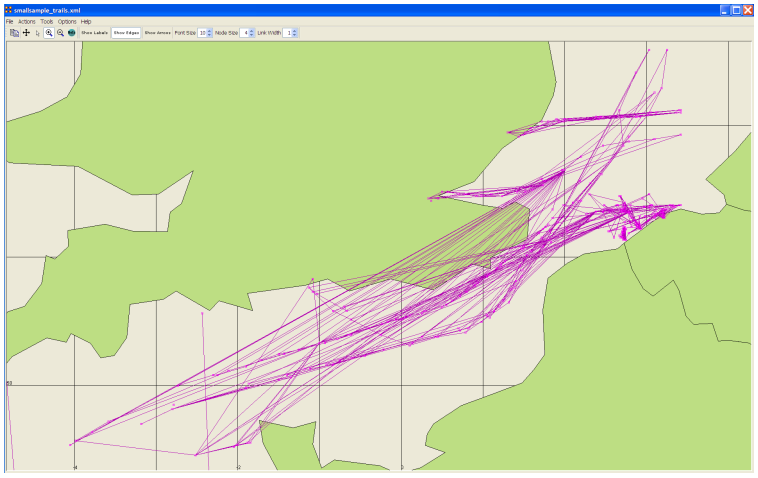

Figure 12: Initial OraGIS analysis

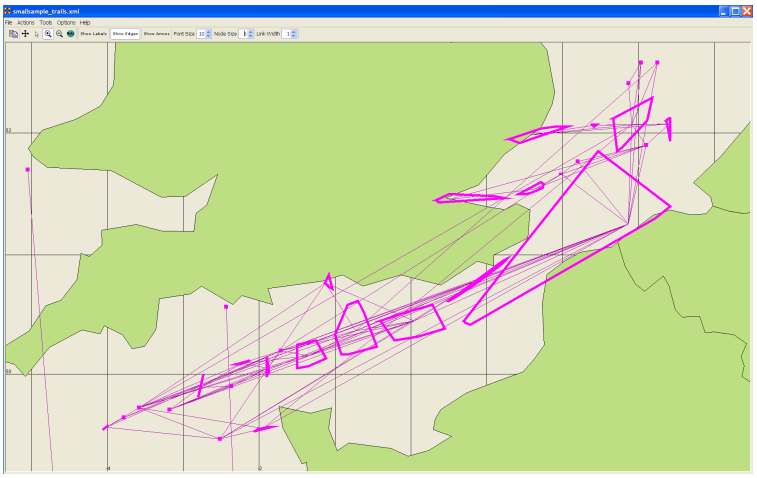

Figure 13: Clustered OraGIS analysis

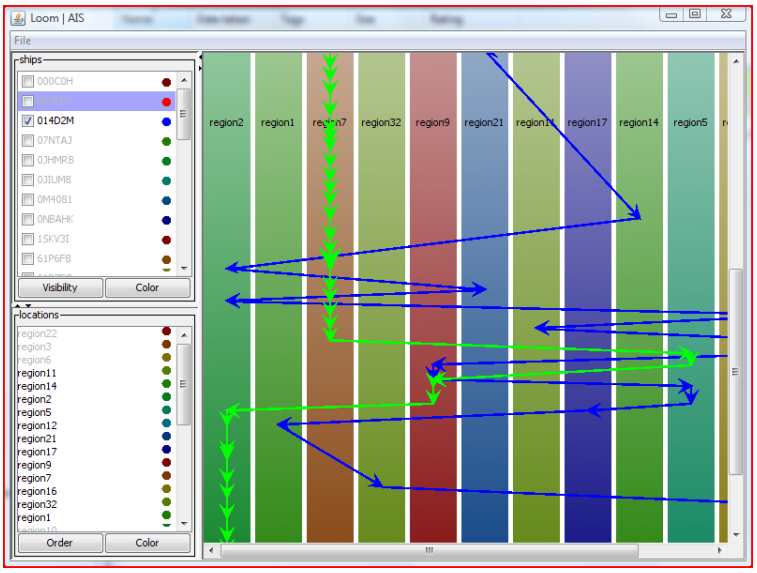

Figure 14: Loom analysis after clustering

matrix yields the graph in Figure 15.

- A ”transition” matrix which connects ports only if a ship traveled directly between them. Visualizing this matrix yields the graph in Figure 16.

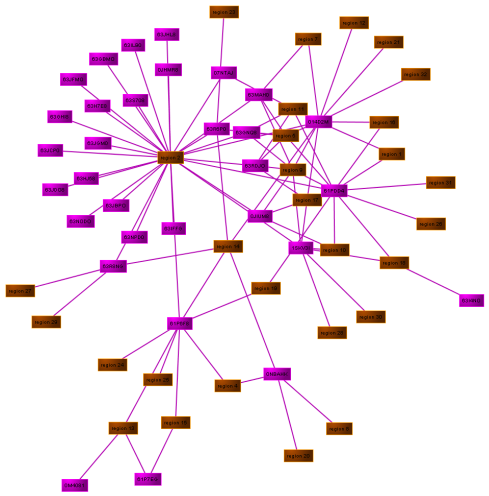

Figure 15: Visitation Network, Ship (Purple) X Region (Orange)

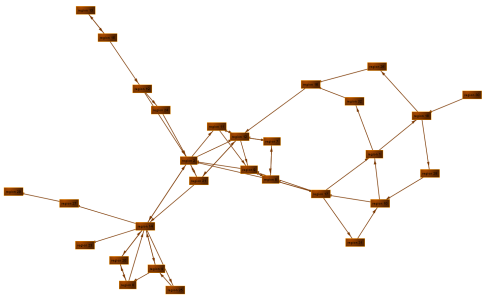

Figure 16: Transition Network, Region X Region

With our data now available in relational, temporal, and spatial forms, we can apply all of our tools to investigating questions. For example, an economist might be interested in the hypothesis that the best connected locations are also those on the shortest routes between other locations. Figure 17 investigates this by graphing the betweenness centrality of locations (their tendency to appear on short paths on short paths) against their eigenvalue centrality (summarizing the degree of connectivity of both them and several degrees of neighbors).

The high degree of correlation suggests that our hypothesis is correct. However, it may be interesting to examine nodes that are significantly above or below the line, indicating that they disproportionately exhibit one form of centrality. Figure 18 puts these results in geographic context by giving a GIS view of the transition matrix with nodes sized according to betweenness centrality.

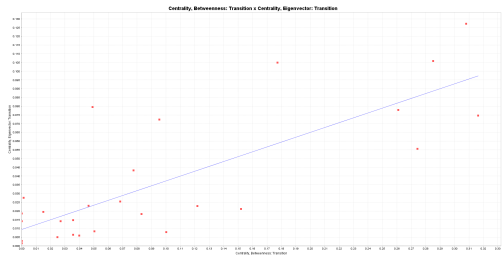

Figure 17: Comparison of centrality measures for nodes

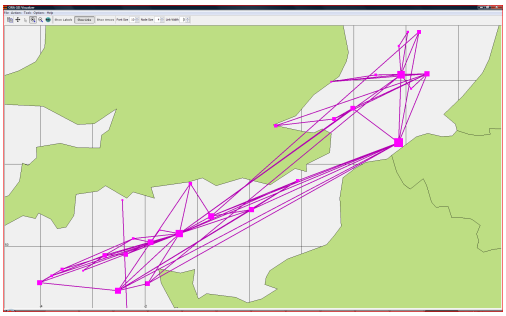

Figure 18: GIS view of transition network, nodes Sized by betweenness

This brief analysis was meant to simply demonstrate how the tools can be chained in practice. For a more detailed examination of this dataset please see [3].

5 Conclusion and Future Work

The ORA suite incorporates new tools for the analysis of networks that exist in space and time. OraGIS allows analysts to visualize their data geospatially and aids them in selecting an appropriate level of aggregation for their target problem. Loom allows subjects to be tracked over time to search for collocated individuals. Loom and OraGIS can be paired to provide a unified platform for the relational analysis of spatial and temporal analysis.

Loom and OraGIS store data through extensions of the DyNetML file format. OraGIS requires geospatial attributes for "Location" nodes, while Loom requires additional temporal information. Both Loom and OraGIS provide graphical user interfaces for the visualization of data. OraGIS emphasizes spatial relations presents network information overlaid on a world map. Loom focuses on temporal information, displaying a waterfall diagram of subjects and their locations. These two tools can be combined to analyze complex real world data, such as AIS records of ship locations.

Future work includes path analysis and prediction through a unified probabilistic framework. This would allow for inference that leverages the interactions between space and time rather treating each factor independently. Proposed work also includes better integration of the information loss metric into the analysis and visualization tools.

Loom and OraGIS allow analysts to use Ora's powerful relational tools to analyze real world

temporal and geospatial data.

References

[1] Kathleen Carley. Dynamic network analysis. In *Committee on Human Factors*, pages 133–145. National Research Council, 2004.

[2] Kathleen M. Carley and Jefferey Reminga. ORA: Organizational Risk Analyzer. Technical Report CMU-ISRI-04-106, Institute for Software Research International, Carnegie Mellon University, 2004.

[3] George B. Davis and Kathleen M. Carley. Computational Analysis of Merchant Marine Global Positioning Data. Technical Report CMU-ISRI-07-109, Institute for Software Research, Carnegie Mellon University, 2007.

[4] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[5] Kathleen Carley Max Tsvekovat and Jefferey Reminga. DyNetML: Interchange Format for Rich Social Network Data. Technical Report CMU-ISRI-04-105, Institute for Software Research International, Carnegie Mellon University, 2004.

[6] National Geospatial-Intelligence Agency. *DMA TECHNICAL MANUAL 8358.1: Datums, Ellipsoids, Grids and Grid Reference Systems*.

[7] Jamie F Olson and Kathleem M Carley. Summarization and Information Loss in Network Analysis. In *Workshop on Link Analysis, Counter-terrorism, and Security held in conjunction with the SIAM International Conference on Data Mining (SDM)*, April 2008.

[8] S Openshaw and S Alvandies. *Geographical information systems : principles, techniques, management, and applications*, volume 1, chapter Applying geocomputation to the analysis of spatial distributions. John Wiley & Sons, New York, 2 edition, 1999.

[9] S Wasserman and K Faust. *Social Network Analysis*, chapter Social Network Analysis and the Behavioral Sciences. Cambridge University Press, New York, 1 edition, 1994.